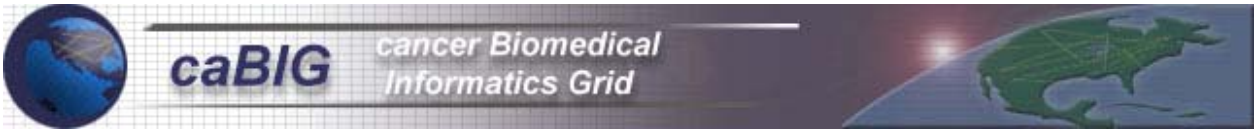

Integrative Cancer Research Special Interest Group Teleconference

Data Analysis and Statistical Methods SIG Meeting Minutes

Date, Time & Location:	May 7, 2004 2:00 – 3:00 EDT
Attendees:	Patrick McConnell – Duke (funded developer) Don Baldwin – Penn (funded adopter) David Fenstermacher (funded adopter) Robert Clarke – Georgetown (funded developer) Joseph Wang – Georgetown (funded developer) Jason Xuan – Georgetown (funded developer) Ajay Jain – UC San Francisco (funded developer) David Jewell – Dartmouth (funded developer) Simon Lin – Duke (funded developer) Jim Lyons-Weiler – Pittsburgh (unfunded developer/adopter) Tom Moloshok – Fox Chase (funded developer) Judith Goldberg – New York (funded adopter) Joe Parker – Lineberger (funded developer) Don Xiang – Lineberger (funded developer) Louise Showe – Wistar (funded adopter, unfunded developer) Lianhong Tang – Vanderbilt (funded participant) Edith Zang – Institute for Cancer Prevention (unfunded adopter) Naveen Vinukanda – Institute for Cancer Prevention (unfunded adopter) Subha Madhavan NCICB Mervi Heiskanen – NCICB Bryan Pittman - CCR Claire Zhu – BAH Juli Klemm - BAH
Introduction:	**Roll-call, open meeting, review meeting goals** - Establish goals and priorities for this SIG - For match making purposes, Developers and Adopters will be asked to give a brief statement of their capabilities and interests, respectively - Identify and define additional research
Overview Discussion:	**Review goals and objectives of Data Analysis and Statistical Methods SIG** Scope: Tools and systems that pertain to sophisticated data analysis. **Open Discussion** - A number of important issues and challenges with regard to experimental scientists' struggle with data analysis were discussed: o There is a large amount of data being generated from complex experiments and most experimentalists are insufficiently equipped to handle the data. o There is a short supply of biostatisticians which makes it difficult to have such a resource available at all desired times in a project. o Many of the currently available tools assume a strong statistical/mathematical background and therefore are difficult to use by experimentalists or, worse, can easily be used incorrectly and produce

misleading results.

- Various ways of addressing this issue were discussed:
  - Provide better training of experimentalists with regard to basic statistics. Coordinate/collaborate with Training WS/WG.
  - Provide better documentation for software with high quality sample data sets. Link to on-line tutorials whenever possible.
  - Encourage experimentalists to be involved at Adopter sites
  - Generally improve communication between biologists and statisticians.

High-level review of Center interests in this SIG

Developers

Dartmouth: Q5 is a tool for analyzing mass spec data, toward the discovery of cancer diagnostics. Currently written for MatLab. For caBIG, would port to Java.

Georgetown: VISDA is a tool for the analysis of microarray data. It is aimed at cluster modeling, cluster discovery and cluster visualization.

UC San Francisco: Magellan allows uploading of various data types including experimental and annotation data; allows statistical analysis and visualization in the context of annotation. For caBIG, would integrate with caArray for microarray data and caBIO for annotations.

Lineberger: DWD is capable of combining microarray data sets from different platforms and different experiments; allows data manipulation across multiple datasets; provides independent validation. Currently in MatLab. Transitioning to JAVA.

Fox Chase: FDGP: currently in-house use; allows image upload, tools for normalization, clustering analysis. Allows multiple, simultaneous data analysis.

Wistar: Algorithms (penalized discriminant analysis) for front-end (background subtraction, normalization, etc.) as well as back-end (bootstrapping, classification) analyses of microarray data.

Pittsburg: caGEDA: http://bioinformatics.upmc.edu/GE2/GEDA.html Allow uploading data with or without accession numbers; on-line tools for normalization, SAM, chromosome mapping etc. Search features includes comparison of differentially expressed genes to pre-loaded literature of single gene and single protein studies reporting differential expression in cancers distilled from >2,600 manually selected abstracts. Special option for permutation analysis; heat-map grid for expression pattern clustering. Available on-line. Paper coming out soon. Also involved in the caProteo initiative, which involves the EDRN and a number of caBIG participants.

Duke: JavaR would provide an interface between Java and R. Could be run in a single-user environment or could be centralized on a high-performance server. RProteomics is a tool that carries out baseline correction and peak finding for MALDI-TOF data. Also aids in biomarker identification. Will be capable of analyzing metabolomic data as well in the future.

Adopters

NYU: Interested in training and implementation issues.

Penn: Interested in many of the tools including VISDA, Magellan, DWD, RProteomics.

Integrative Cancer Research Special Interest Group Teleconference

	Wistar: Interested in FDGP and Magellan. Institute for Cancer Research: Interested in Magellan, caGEDA, VISDA.
Other discussion items:	**Mission Statement** Ajay Jain from UC San Francisco will draft a Mission Statement for this SIG. This draft will be circulated to the group for comments. **Regular meeting schedule** The group will meet on the first Friday of each month at 2:00 Eastern. Ongoing frequency will be revisited if necessary. **Notes and additional information** Meeting minutes will be emailed and posted on the caBIG on-line forum after review by the group: http://ncicbforums.nci.nih.gov/forums/cabigforum/lfs/icrlfs/SIGs/data_analysis_stats Developers should also post additional information about their tools here.

Action Items:

Name Responsible	Action Item	Date Due	Notes
Ajay Jain	Draft Data Analysis and Statistical Methods SIG mission statement	5/28/04 (approx)	
Juli Klemm	Contact info for this SIG	5/17/04	Will wait for those who do not wish to be included to identify themselves.
Juli Klemm	Distribute meeting minutes	5/17/04	
Juli Klemm	Schedule ongoing meetings	5/17/04	
All developers	Post relevant additional information to the caBIG on-line forum (see link above)	5/14/04	